An Augmented Reality Setup with an Omnidirectional Camera based on Multiple Object Detection

Tomoki Hayashi, Hideaki Uchiyama, Julien Pilet and Hideo Saito
Keio University
3-14-1 Hiyoshi, Kohoku-ku
Yokohama, Japan
{*tommy,uchiyama,julien,saito*}*@hvrl.ics.keio.ac.jp*

***Abstract*—We propose a novel augmented reality (AR) setup with an omnidirectional camera on a table top display. The table acts as a mirror on which real playing cards appear augmented with virtual elements. The omnidirectional camera captures and recognizes its surrounding based on a feature based image retrieval approach which achieves fast and scalable registration. It allows our system to superimpose virtual visual effects to the omnidirectional camera image.**

In our AR card game, users sit around a table top display and show a card to the other players. The system recognizes it and augments it with virtual elements in the omnidirectional image acting as a mirror. While playing the game, the users can interact with each other directly and through the display. Our setup is a new, simple, and natural approach to augmented reality. It opens new doors to traditional card games.

***Keywords*-augumented reality; multiple object detection; bag-of-features; omnidirectional camera;**

I. Introduction

Recent progress in computer vision significantly extended the possibilities of augmented reality, a field that is quickly gaining popularity. However, augmented reality is young and few example of well designed and useful applications exist. More researches in the domains of AR interfaces and interaction methods will be required to reach maturity [1].

In this context, we propose a new AR setup that allows multiple users to naturally interact with each others, both directly and through an augmented reality system. Our setup relies on an omnidirectional camera lying on a horizontal screen acting as a table around which users sit. The screen displays the camera output, making it look as a mirror.

Our setup has several interesting properties. A mirror is a natural and friendly object that will not scare any user, as opposed to head mounted display for example. Because the setup remains simple and natural, it is user friendly and easy to understand. The system augments reality with little intrusion. The contact with the other players remain real, the AR technology is not acting as a barrier.

From a technical point of view, our setup is based on an image retrieval system. In our context, we need to retrieve quickly about 100 targets appearing on an omnidirectional image. For the retrieval part, we rely on Nister's vocabulary tree to quantize SIFT features [2], [3].

As an application, we propose an AR card game system. Users play a card game around our magic table. When they show a card to other players, the system also recognizes it. The reflections of the card on the mirror is augmented with some virtual content related to the game. Because the system is able to augment several cards at the same time, it can show interactions. For example, a player's card can throw a fireball to the card of his opponent. Compared to previous AR card games [4], [5], our game system is simple and efficient in terms of physical configuration. Moreover, our setup is not limited to two players.

II. Related works

First, we review image retrieval methods related to our system. Recently, local keypoint descriptor based approaches are a mainstream trend for image retrieval. SIFT [3] and SURF [6] are robust to a camera rotation and translation, and illumination change. Since these descriptors are high dimensional vectors such as 128 dimensions with larger computational costs, Sivic and Zisserman [7] quantized the SIFT descriptors by a single depth based k-means tree to achieve run-time object retrieval throughout a movie database. These quantized descriptors are usually called *visual words*, which are generated by a vocabulary tree [2], [8], [9], [10]. Visual words are collected to depict an object and exploited to a classification of objects by ignoring a position of each visual word [11]. Nister and Stewénius have proposed a hierarchical k-means tree as a tree structure [2]. Since they reported highly scalability and high speed performance, we apply this method to multiple image retrieval.

Omnidirectional cameras have many applications, because they can capture 360 degrees images. Wallhoff *et al.* [12] proposed face tracking in a meeting room. Douxchamps and Campbell [13] presented a real-time face detection, tracking and characterization from omnidirectional video. Other applications by setting the omnidirectional camera on a table capture people sitting around without uncomfortable feeling [14], [15].

Our key idea is to achieve the omnidirectional augmented reality system by setting the omnidirectional camera on a display lying on the table. Users can see themselves on the

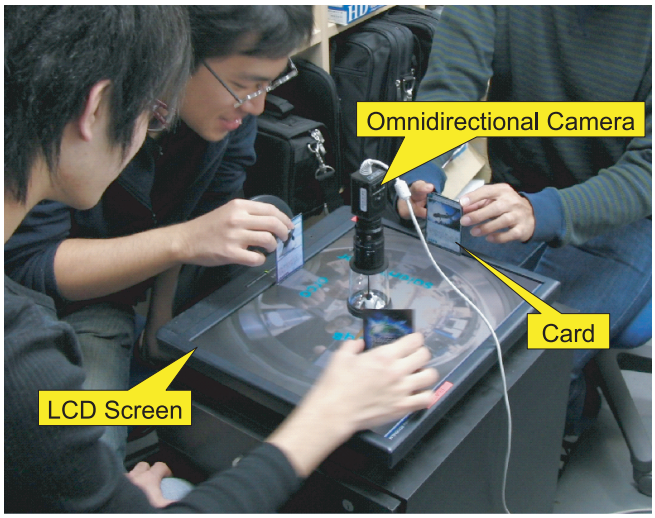

Figure 1. Configuration of the proposed system. An omnidirectional camera is set on a standard LCD screen. If some cards are shown to the camera, our system recognizes it and annotations or augmentation is displayed on the LCD screen as an augmented reality.

screen, as in a mirror, and the mirror image is augmented when the system detects objects around the camera. By integrating a framework of multiple object retrieval, the augmentation and interaction can be achieved.

III. Proposed system

The configuration of our system is made of two components; an omnidirectional camera based on a hyperbolic mirror for capturing the environment, and a screen to display the captured image. Fig. 1 depicts our system. The camera is centered on the display, and calibrated for clearly capturing its surrounding. Since the center of the display corresponds to the vacant region of the omnidirectional image, setting the camera on the center does not break the continuity of the displayed image. The display can be a standard LCD screen or a projector fixed above or under the table.

In the system, users play a card game and show some of them to other players. The camera also observes the card, and our system detects it quickly and locates it precisely. The detected cards are shown with some annotations associated with their content and status in the game. The augmented image is displayed naturally, as a reflection of the reality.

A. Building the card database

As a pre-processing, the target cards are indexed in a database for multiple object retrieval.

The cards are captured by the omnidirectional camera set in front of them. The omnidirectional image is geometrically rectified into a panoramic image to better match the SIFT assumptions. Even though the panoramic image has a cylindrical projection instead of a planar one, SIFT works well because a planar projection is a good local approximation of the cylindrical one.

To add a new card in the database, we manually segment it from an omnidirectional image, by clicking the upper-left and lower-right corners. After segmenting a card region, the card can be virtually annotated. In addition, the relationship between several objects can be also registered in order to achieve augmentation showing interactions between cards.

When the system is running, after the retrieval process, annotations of visible objects are overlaid at the appropriate position.

B. Image Retrieval

To quickly determine which card is visible and where, we use the well known SIFT features [3]. Basically, the system has a picture of each card it is supposed to recognize. Features are extracted from this picture. The SIFT descriptors have 128 dimensions, which is precise but inconvenient for indexing. Therefore, we vector quantize them. To do so, we first collect a large number of descriptors. We then cluster them using recursive K-mean, with $K = 4$ over 8 levels [2]. The resulting tree has about 60000 leafs. At runtime, given a 128 dimensional descriptor, our system can go down the tree, comparing cluster centers at each node, to find the leaf the descriptor belongs to.

Such a vector quantization improves indexing efficiency: each leaf of the tree has a table pointing to the indexed cards on which a matching descriptor appears. Retrieval is then quick, because looking up the tables associated with the features found on the query quickly yields a list of candidates. The candidates are then ranked, and the best ones go through a verification process.

To explain the ranking score, let Q be a query image, and q_i the number of features detected in Q with a descriptor quantized to the leaf i. Similarly, let A denote a card in the database. The score between Q and A is:

$$S(q,a) = \frac{1}{\sum_i a_i} \sum_i \min(q_i, 1) \min(a_i, 1) \ .$$

We consider all the database entries with a score above 0.05 for verification.

To verify if a candidate is actually visible or not, we match in turn the full SIFT descriptors of each candidate, as described in Lowe's paper [3]. If at least 3 can be reliably found, we consider the object as detected. The location of the card is determined to be at the average location of the matched points.

C. Augmentation

When objects are detected, the corresponding annotations are overlaid on the omnidirectional image. The type of the annotation includes, but is not limited to words, animations, pictures, computer graphics and so on. Since the objects are localized, the annotations are drawn at their appropriate

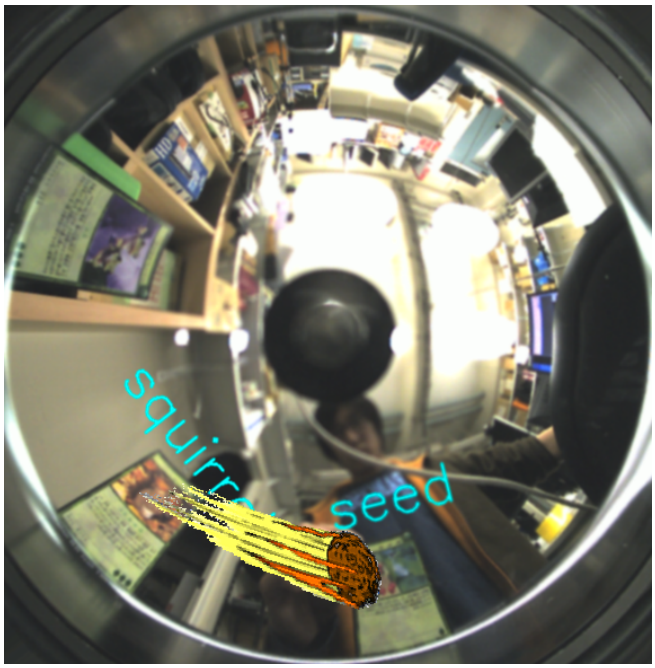

Figure 2. Augmented omnidirectional image. Displayed cards fight each other. A fighting result is shown as some augmentations. In this case, since the squirrel is stronger than the seed, the former shoots the latter a fire ball.

position in the omnidirectional image. In addition, some relationships between objects can be also considered. If a set of interacting objects is detected, the system can display annotations or augmentation visualizing their relationship as depicted by Fig. 2. In the example of a card game, the strength of each card is registered in the database and the system can determine which played card wins. The score of each user can be also computed by the system automatically based on the rule of the game. The information about which player wins and on the sore of each player is also displayed. With such a setup, players of the card game can play naturally without knowing detailed rules.

IV. Experiments

The omnidirectional camera we used is a digital camera combined with a hyperbolical mirror. The original image resolution is 1376×1376 pixel. We rectify it to a panoramic image of 1024×145 pixel. The vertical resolution of the panoramic image is limited, because the image region in which the object may appear is limited. Such a low resolution greatly reduces computation time. We registered 60 cards (My Earth Projects LLC published in 2008 [16]) in the system for our experiments.

In a first experiment, we evaluate our system's ability to retrieve and locate multiple objects in a panoramic sequence. In this experiment, randomly selected cards are shown to the camera. In the case depicted by Fig. 3, all cards were detected at the correct location and objects are annotated. Manual testing of the setup shows that the cards are detected if they are facing the camera and if they are close enough. If the distance between the card and the camera is more than about 30 cm, it can not be detected, because the card becomes too small on the image. The SIFT matching process handles in-plane rotation well (illustrated in Fig. 4), and a limited amount of perspective effects. Overall, the system is robust enough for our purpose.

The processing time also has been computed as table I.

The resolution of our omnidirectional camera causes a long acquisition time. To mitigate this effect, we parallelized the capture and the processing. Image retrieval is fast thanks to the inverted tables. Because extracting SIFT features is an computationally heavy operation, we used the GPU implementation of [17]. As a result, the entire process takes 260 msec, which is enough for our interactive purpose.

To evaluate the quality of our system's interface, we conducted the following experiment. Three people sit around the table and play the card game, enjoying the augmentation displayed by our system. When the users showed cards to other players, their reflections on the screen were augmented with some annotations. Stronger cards could even throw a fireball to weaker ones. Because this augmented reality system is unobtrusive and let users interact directly, playing feels natural.

V. Conclusions and Future Works

We have proposed a novel AR setup showing augmented content on real cards while allowing users to naturally interact with one another. The system is based on an efficient image retrieval method and uses an omnidirectional camera. The users play a card game around a magic table that shows augmented reflections. When a player shows a card to the others, the system recognizes which card it is. If it is stored in a database, the corresponding annotations are overlaid on the omnidirectional image and displayed on the table. Furthermore, our method is able to detect multiple objects simultaneously, making it possible to produce virtual interactions between real cards.

Acknowledgment

This work is supported in part by a Grant-in-Aid for the GCOE for high-Level Global Cooperation for Leading-Edge Platform on Access Spaces from the Ministry of Education,

Table I
PROCESSING TIME

task	**computing time (msec)**
image capture	256
feature extraction	83.5
object retrieval	6.20
descriptor matching	51.0

Figure 3. A result of multiple object detection. The object detection scheme is achieved in a panoramic image transformed from a omnidirectional image. In this image, four cards are shown in front of the camera, and all cards are recognized successfully. We should take note of that the annotations and the panoramic image are not displayed on the screen but shown here as just detection results.

Figure 4. A result of a robust detection of the card. In this image, though the left card is rotated drastically, our system recognizes it by applying the SIFT matching to judge whether candidate cards exist in the image or not.

Culture, Sport, Science, and Technology in Japan and Grant-in-Aid for JSPS Fellows.

REFERENCES

[1] C. Scherrer, J. Pilet, V. Lepetit, and P. Fua, "Souvenirs du monde des montagnes," *SIGGRAPH*, vol. 42, no. 4, pp. 350–355, 2009.

[2] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. CVPR*, 2006, pp. 2161–2168.

[3] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[4] A. H. T. Lam, K. C. H. Chow, E. H. H. Yau, and M. R. Lyu, "ART: augmented reality table for interactive trading card game," in *Proc. VRCIA*, 2006, pp. 357–360.

[5] W. Lee, W. Woo, and J. Lee, "Tarboard: Tangible augmented reality system for table-top game environment," in *Proc. IWPG*, vol. 5, 2005.

[6] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," *CVIU*, vol. 110, pp. 346–359, 2008.

[7] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. ICCV*, 2003, pp. 1470–1477.

[8] H. Jégou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *IJCV*, pp. 1–21, 2009.

[9] F. Jurie and B. Triggs, "Creating efficient codebooks for visual recognition," in *Proc. ICCV*, vol. 1, 2005, pp. 604–610.

[10] F. Li, W. Tong, R. Jin, A. Jain, and J. Lee, "An efficient key point quantization algorithm for large scale image retrieval," in *Proc. LSMMRM*, 2009, pp. 89–96.

[11] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. ECCV*, vol. 1. Citeseer, 2004, p. 22.

[12] F. Wallhoff, M. Zobl, G. Rigoll, and I. Potucek, "Face tracking in meeting room scenarios using omnidirectional views," in *Proc. ICPR*, 2004.

[13] D. Douxchamps and N. Campbell, "Robust real time face tracking for the analysis of human behaviour," in *Proc. MLMI*, 2007, pp. 1–10.

[14] V. Rozgic, K. Han, P. Georgiou, and S. Narayanan, "Multi-modal speaker segmentation in presence of overlapped speech segments," in *Proc. Multimedia-Volume*, 2008, pp. 679–684.

[15] R. Stiefelhagen, X. Chen, and J. Yang, "Capturing interactions in meetings with omnidirectional cameras," *IJDET*, vol. 3, no. 3, pp. 34–47, 2005.

[16] My Earth. http://myearth.ne.jp/.

[17] S. Sinha, J. Frahm, M. Pollefeys, and Y. Genc, "Gpu-based video feature tracking and matching," in *Proc. EDGE*, vol. 278, 2006.